ElectroSphere: Electrical Electronics Engineering Bulletin

PRIME OPEN ACCESS

Volume 1, Issue 2

Research Article

Date of Submission: 14 June, 2025
Date of Acceptance: 06 Aug, 2025
Date of Publication: 12 Aug, 2025

The Role of AI in Mental Health Crisis Management

Sanjay Kumar*

Independent Researcher, India

***Corresponding Author:**
Sanjay Kumar, Independent Researcher, India.

Citation: Kumar, S. (2025). The Role of AI in Mental Health Crisis Management. *Electro Sphere Electr Electron-ic Eng Bull*, *1*(2), 01-07.

Abstract
The increasing incidence of mental illness, particularly among youth, has raised important gaps in early intervention and access to assistance. With so many people seeking online mediums to convey emotional turmoil, there is the possibility of understanding how Artificial Intelligence (AI) might be used proactively to identify and address mental health emergencies. This paper examines the potential of AI systems in identifying indicators of self-harm, suicidal thoughts, and psychological distress via natural language chats. Through speech patterns, sentiment, and behavioral indicators analysis, AI-powered tools can be used as precursors that can escalate issues to the potential experts or emergency services. While the promise of such technology is great, the application gives rise to difficult questions regarding consent, data privacy, false alarms, and ethical accountability. By examining current technologies, theoretical models, and intended intervention designs, this paper seeks to establish a foundation for creating responsible, effective, and empathetic AI systems that not only recognize human emotion but act when intervention is most urgently needed.

Introduction
MEntal illness is a rapidly increasing public health issue. Conditions such as depression, anxiety, and suicidal thoughts cut across all ages and socioeconomic backgrounds. While there has been growing awareness and policy initiatives to expand mental healthcare access, a large number of people suffering from psychological distress go undiagnosed and untreated. Perhaps the largest drawback is lack of ability of early identification of those who are at risk. According to the World Health Organization (2021), over 60% of people with mental health conditions do not seek professional help, often due to stigma and insufficient access to care.

As web-based support platforms and digital communication devices become omnipresent, more and more people now convey their emotional state in text-based communications. This shift in preference opens up an opportunity to leverage the advancements in Artificial Intelligence (AI), and more so in Natural Language Processing (NLP), to identify patterns that indicate mental health crises. With proper training, AI systems can examine language, tone, and context in realtime to recognize early warning signs of self-injury, suicidal ideation, or extreme psychological distress.

This paper explores the potential for AI to enhance mental health crisis intervention systems. More precisely, it examines the feasibility of using AI-based models to identify high-risk behaviors through dialogue and to trigger tiered intervention protocols—spanning from empathic engagement to emergency escalation, including potential alerting of emergency services like 911. It also considers the ethical, legal, and technical dimensions of such interventions, such as consent of the user, data protection, algorithmic efficacy, and potentialities for abuse or harm. By the incorporation of insights from past mental health technologies, AI ethics studies, and crisis intervention models, this paper aims to contribute to the development of responsible and effective AI systems that enable early intervention, reduce unnecessary harm, and promote human-delivered mental health care.

Background and Literature Review
The intersection of artificial intelligence (AI) and mental health services has opened up new possibilities for the

enhancement of early detection, care, and intervention in the context of psychological crises. Particularly, the use of AI systems for the detection of suicidal thoughts in real-time and the activation of related emergency interventions is a pressing, albeit understudied, use of the technology.

Early uses of AI in mental health were in diagnosis, where machine learning was applied to identify mental health conditions from physiological and behavioral data. In a systematic review of more than 200 studies, A. B. Shatte, D. M. Hutchinson, and S. J. Teagu in their research indicated that AI models were moderately to highly accurate in identifying conditions like depression, PTSD, and anxiety, particularly with multimodal inputs of data [1]. These models, however, functioned mostly in static or retrospective environments, providing minimal support for real-time intervention.

Current studies have explored the ability of natural language processing (NLP) in suicidal ideation identification. J. P. Pestian demonstrated that supervised machine learning models have the ability to identify suicidal and non-suicidal language in adolescents accurately based on linguistic features [2]. Similarly, G. Coppersmith, R. Leary, P. Crutchley, and A. Fine analyzed social media posts to identify individuals at high suicide risk by extracting patterns of language that reflect hopelessness and social isolation [3]. While these studies have promising results, they are primarily concerned with risk detection and lack escalation or emergency response processes.

Chatbots like Woebot, Wysa, and Tess have increasingly been found to provide cognitive behavioral therapy (CBT) and emotional support through chat interfaces. Specifically, Woebot has been found to be effective in decreasing depressive symptoms in young adults over short time frames [4]. However, it is important to mention that these bots are generally meant for general wellness support and lack the capability to identify or address acute mental health crises.

The ethical concerns surrounding the application of artificial intelligence within critical mental health situations have been extensively discussed. E. Vayena, A. Blasimme, and I. G. Cohen contend that algorithmic systems applied in healthcare need to be regulated by strong ethical frameworks, especially where such systems have the potential to violate the confidentiality of users in an attempt to alert emergency services [5]. The balance between safety and privacy is particularly emphasized where an AI concludes that a user is in a critical risk of self-harm. Despite these advancements, there are important gaps in the academic literature.

The majority of the prevailing models are either primarily diagnostic or facilitative but lack an actual real-time feedback mechanism to enable intervention where acute psychological distress arises. In addition, contemporary systems are frequently susceptible to suboptimal performance in noncontrolled settings, and therefore there is a necessity for adaptive models that can appreciate context, culture, and variations between individuals. There is also a critical lack of interdisciplinary effort between technologists, mental health practitioners, and legal practitioners in the development of systems that are both efficient and ethically justifiable.
This work seeks to close these gaps by suggesting an AI-driven system capable of detecting high-risk patterns of behavior from real-time dialogue and initiating tiered responses—e.g., calling authorities such as emergency services—depending on the level of detected risk.

Methodology

This paper proposes a conceptual framework for integrating real-time AI-based mental health risk detection with actionable user interface (UI) and system-level interventions. Unlike existing research that focuses primarily on the classification of mental health status from user input, this study seeks to bridge the operational gap between risk detection and escalation to human or institutional support. The methodology is design-focused, with a strong emphasis on feasibility, ethical deployment, and extensibility in existing digital platforms.

Overall System Design

The proposed architecture is composed of three interconnected layers:

- **Detection Layer:** Leverages state-of-the-art AI models to assess user mental health risk in real time.
- **Interface Layer:** Provides conversational, contextsensitive, and ethical engagement with the user.
- **Intervention Layer:** Implements a structured response protocol based on risk level, including escalation to human professionals or emergency services.

Each layer is discussed in depth below.

Detection Layer: AI-Based Risk Assessment

The detection layer employs pre-existing transformer-based NLP models trained to identify markers of depression, selfharm ideation, and suicidal thoughts. These models may be sourced from open-access repositories (e.g., CLPsych datasets, Reddit Self-reported Depression Dataset) or deployed via commercial APIs that provide mental health sentiment scoring. Models such as:

- BERT fine-tuned for suicidality classification,
- RoBERTa-based multi-class emotion detection, and
- MentalBERT have demonstrated high accuracy in recognizing suicidal ideation from text-based inputs. These models generate confidence scores or probabilities associated with a spectrum of emotional and psychological states. Key aspects:

- **Continuous Monitoring:** The system monitors each conversational turn, with cumulative risk scoring over a session.
- **Threshold Triggers:** A dynamically adjustable threshold triggers an ”at-risk” status, which is then handed off to the Interface Layer.

Interface Layer: Real-Time Conversational UI Interventions
This layer ensures ethical and humane engagement when the AI model flags high-risk behavior. Drawing from OpenAI’s Function Calling paradigm and similar plugin-based architectures, the system integrates intervention modules that can be invoked contextually by the AI.
Key components:

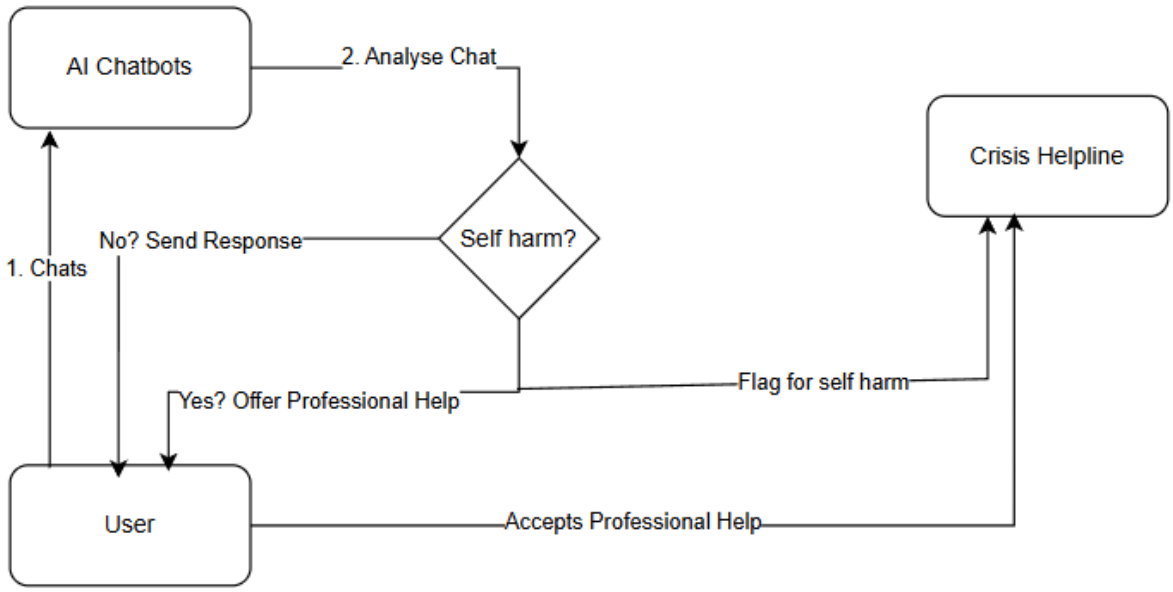
AI-Initiated Dialogues: When a high or critical risk level is detected, the system initiates a soft check-in with the user: “I’m here for you. It seems like you’re going through a really difficult time. Would you like to speak to someone right now?”

- **Actionable Prompts:** Users are presented with immediate, actionable choices:
- “Yes, connect me to a professional”
- “I just want to talk a bit more”
- “I’m okay for now”
- **User Consent Gateway:** At all stages, escalation actions are gated by user consent wherever possible, unless there is an imminent danger of harm.
- **Contextual Logging:** All flagged messages and the AI’s decision path are logged locally or to a secure backend (depending on deployment), enabling human review.
- **Function Plugins:** Just as OpenAI plugins allow models to execute real-world functions, the system integrates callable functions such as:
- flagUserToLocalAuthority(userId)
- connectToLiveCounselor(userId)
- sendCrisisAlert(location, riskLevel)
- logSessionForReview(sessionId)

These functions operate independently of the AI model, ensuring separation of concerns and modular design.

Intervention Layer: Escalation Protocol and External Connectivity
Once a user is flagged as at-risk or critical, the system routes the interaction through a structured, multi-tiered intervention pipeline:

- Risk Level Moderate: Encourage ongoing conversation; offer professional help voluntarily
- Risk level High: Immediate AI check-in + present counselor options + optional logging
- Risk Level critical: Alert moderators/clinical team + initiate emergency protocol (with opt-out only in nonlife-threatening situations) Additional features include:
- Professional Network Integration: Integration with mental health platforms (e.g., BetterHelp, iCall, Talkspace) for instant connection.
- Local Authority APIs: Where available, APIs for national or regional emergency services (e.g., 911 in the U.S., 112 in Europe, iCall in India) can be triggered with user consent or automatically if legally allowed.
- Geo-fencing Logic: Determines which local authority to alert based on user IP or declared location.



Data Governance and Privacy Mechanisms
As mental health data is highly sensitive, the system incorporates a privacy-first design:
- No data storage by default unless explicit consent is provided.
- Differential privacy and end-to-end encryption where data logging is required.
- Audit trails for every AI-triggered escalation to ensure accountability.

Future Implementation and Validation
While the present study does not perform empirical testing, future work includes:
- Deploying a prototype in a closed trial with trained human moderators.
- A/B testing user outcomes with and without active escalation logic.
- Survey-based validation from mental health professionals regarding the ethics, sensitivity, and effectiveness of conversational strategies.

This methodology prioritizes ethical responsibility, real-world feasibility, and modular integration with existing AI platforms. By introducing a clear escalation pipeline triggered by proven NLP techniques and embedded into conversational interfaces, the proposed system aims to bridge the critical last mile in digital mental health intervention.

Ethical and Legal Considerations

The application of artificial intelligence (AI) in responding to mental health crises is riddled with significant ethical and legal complexities. Despite the goal of safeguarding and assisting individuals who may be at risk of self-harm, the functioning of such systems must contend with the problems of consent, privacy of information, informed intervention, algorithmic bias, and legal responsibility. The following section describes the most significant ethical and legal frameworks that govern the design and implementation of the suggested system.

Informed Consent and User Autonomy One of the key ethical principles of any AI-assisted intervention in mental health is the principle of informed consent. Users of an AI must be clearly informed—at the beginning—of the capabilities of the system in terms of identifying psychological distress and initiating interventions. Furthermore, the system must offer users options to opt in or opt out, especially in the case of non-critical levels of concern, thereby allowing users to have control over the use of their data and over decisions regarding any escalation.

Where an individual's life is threatened, the question of overriding autonomy is a serious one. Global mental health guidelines suggest that confidentiality violations are acceptable where an individual is in immediate danger of harming others or themselves. Nevertheless, even there, intervention must be consistent with established legal procedures and ethical rationales.

Data Privacy and Confidentiality Given the very personal nature of issues involved in mental health discussions, maintaining data integrity is of paramount importance. The system has to comply with privacy regulations that are jurisdiction-specific, such as the General Data Protection Regulation (GDPR) in the European Union, the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and India's Digital Personal Data Protection Act (DPDP Act). Major privacy safeguards are:
- End-to-End Encryption of all communications.
- Anonymization or pseudonymization of user identities where feasible.
- Minimal data retention practices, keeping only what is necessary and only with the user's consent.
- User audit trails and access logs to provide system behavior transparency and access.
- These practices are not just necessary for regulatory compliance, but for building confidence with vulnerable consumers.

Ethical Intervention and Escalation While it is technologically feasible to detect a mental health crisis, to respond appropriately is contextdependent and subtle. An AI system cannot—and should not—substitute for professional human judgment, especially in high-risk psychological conditions. The function of the AI is therefore restricted to triage and referral—not diagnosis or treatment.

Three principal ethical protections are recommended:
- **Soft Escalation Pathways:** Users need to be prompted and offered choices prior to initiating contact with mental health workers or emergency services.
- A proportional response requires escalation to match the level of perceived risk to avoid unnecessary alarm or interference.
- **Human-in-the-loop monitoring:** All high-risk sessions indicated should be checked by human monitors or clinical staff prior to making life-changing decisions, where possible.

Algorithmic Fairness and Bias

Language data-trained AI models can inherit biases in training data, leading to unfair or inaccurate classifications across different demographic groups (e.g., socioeconomic status, gender, race). Misclassification in mental health can have serious consequences, from unnecessary distress and false alarms to failure to intervene in a timely manner. To

counteract this:

- Models must be evaluated for fairness using fairness metrics both during training and testing.
- The system must be linguistically and culturally adaptable, particularly in multicultural and multilingual environments.
- Continuing post-deployment surveillance must be used to identify and counteract biased behavior in practice.

Legal Liability and Regulatory Compliance The legal status of AI-driven mental health interventions is unclear. Some of the legal issues include:

- **Liability:** If an AI system does not recognize a crisis or inappropriately escalates, assigning blame—whether with the developer, deployer, or AI service provider—is unclear under the law.
- **Jurisdiction:** In cross-border deployments, there are conflicting data legislation, mental health legislation, and intervention standards that may be problematic.
- **Licensure:** Integration with healthcare practitioners or services can involve clearance as a medical software or device by regulatory agencies such as the FDA (United States) or MDR (European Union).

Until regulatory standards change, developers will need to adhere to the highest ethical standards and seek advice from legal experts to comply.

Psychological Safety and User Experience

Even when used ethically and legally, tone, timing, and language within a system have the potential to influence a vulnerable user's emotional state. Interventions should:

- Empathic and nonjudgmental,
- Carefully crafted to avoid overwhelming the user, and in purpose, especially when data is being shared or external help is contacted. User input should always be gathered in an effort to enhance interaction flows and prevent accidental emotional harm.

Ethical Overview In brief, while artificial intelligence holds immense promise to significantly influence mental health crisis intervention, its application must be built on a firm ethical and legal base. Human dignity, autonomy, and trust are not only prudent to be preserved but must be preserved. Any protection system created must be crafted with diligence itself.

Discussion and Future Work

The integration of artificial intelligence in mental health crisis intervention presents both unprecedented opportunities and complex challenges. This paper has proposed a conceptual framework where AI models embedded in conversational platforms—such as chat-based assistants—are empowered to detect signs of severe psychological distress and initiate structured interventions. While technically feasible, this innovation brings to light a multi-faceted landscape of ethical, technical, operational, and societal considerations.

Bridging the Detection-Intervention Gap

Most existing AI models in the mental health domain, including those trained on social media content or clinical transcripts, primarily focus on diagnostic or predictive capabilities. These models can estimate the likelihood of depressive symptoms, detect suicidal ideation, or categorize anxiety severity. However, few systems effectively bridge the gap between detection and real-time actionable intervention. The proposed framework advances this frontier by embedding escalation mechanisms directly within the user interface, offering timely options such as professional connections, soft nudges, and emergency alerting when necessary.

The novelty lies not in detecting distress alone, but in deploying a user-centric, ethical, and functional escalation architecture that can be modularly integrated into existing platforms—such as ChatGPT—using technologies like OpenAI functions, API calls to verified mental health services, or push notifications to registered caregivers.

Technical Limitations and Infrastructure Challenges

Despite promising strides, the implementation of such systems is contingent upon several unresolved technical challenges:

- **Contextual Understanding:** Current large language models (LLMs) are proficient at recognizing explicit signals of distress but struggle with subtle or culturally nuanced cues. False positives or missed alerts could undermine user trust or safety.
- **Real-Time Responsiveness:** Implementing instant and localized escalation—such as notifying nearby mental health authorities—requires robust back-end infrastructure and integration with local databases, APIs, and networks, which may vary drastically across regions.
- **Scalability:** Deploying such systems globally necessitates scalable multi-lingual, multi-modal support, capable of handling text, voice, and image-based interactions while maintaining high accuracy.
- **Human Oversight:** Though automation is essential for scale, human-in-the-loop designs remain critical for ethical governance. This necessitates building seamless interfaces between AI, moderators, and certified professionals.

Addressing False Positives and Escalation Burnout

Another core challenge lies in managing false positives—cases where users are incorrectly flagged as high-risk.

Over-alerting can result in intervention fatigue for mental health workers, erosion of trust in the platform, and potential distress for users. Future research must focus on adaptive confidence thresholds, cross-validation with historical interaction data, and feedback loops where users can contest or confirm system actions.

Societal and Institutional Readiness

The success of such systems is not merely a technical feat—it is a socio-institutional undertaking. It requires:

- Collaboration between governments, mental health NGOs, AI developers, and legal institutions.
- Establishing cross-border guidelines for AI-aided mental health triage.
- Cultivating public trust through transparent communication and community engagement.

Without institutional readiness and public participation, even the most technically sound systems may face resistance or misuse.

Future Work

To evolve this framework into a deployable and effective tool, the following future research and development paths are proposed:

- **Pilot Studies and Controlled Deployments:** Conducting small-scale pilot implementations in partnership with mental health organizations to test usability, accuracy, and escalation effectiveness in real-world environments.
- **Integration with Mental Health APIs and Services:** Building a library of integrations with existing platforms like iCall, Samaritans, BetterHelp, and 988 Suicide & Crisis Lifeline to allow seamless human-AI collaboration.
- **Feedback-Driven Learning and Personalization:** Developing user feedback loops where the AI adapts its detection algorithms based on historical patterns and user-confirmed signals, enhancing sensitivity and specificity.
- **Cross-Cultural and Linguistic Expansion:** Training models to recognize distress patterns across languages and cultures, avoiding the one-size-fits-all approach that dominates current global AI tools.
- **Ethical Governance Models:** Creating AI ethics committees and governance frameworks to review and approve deployments, ensuring continued alignment with human rights and mental health standards.

Early Intervention Through Educational Integration

A critical frontier for improving mental health outcomes lies in early and proactive engagement. To that end, integrating the proposed AI system within school ecosystems offers a transformative opportunity. Students could be registered on the platform during their early academic years, making them familiar with the tool and encouraging open communication from a young age.

Given that many mental health challenges—including depression, anxiety, academic pressure, bullying, and ragging—originate or intensify during school years, this early exposure can establish trust, normalize seeking help, and create a preventive culture. The interface design must be studentfriendly, with age-appropriate language, gamified well-being prompts, and the ability to escalate concerns silently in environments where verbal expression is difficult.

This strategic embedding within educational systems ensures that students grow up with the knowledge that help is accessible, confidential, and compassionate. It also enables longitudinal data tracking (with full ethical and parental consent) that can offer early warning signs long before a crisis peaks.

AI-driven mental health crisis intervention is not a distant ideal—it is an emergent necessity. The proposed framework emphasizes intervention over observation, human-centered design over algorithmic abstraction, and collaboration over isolation. While challenges remain, this research represents an essential step toward the development of AI systems that not only understand distress but actively and ethically work to alleviate it.

Conclusion

The integration of artificial intelligence into the domain of mental health crisis management presents a profound opportunity to revolutionize early detection and intervention mechanisms. This paper proposed a system capable of analyzing user interactions in real-time, identifying signs of psychological distress—particularly those indicative of selfharm or suicidal ideation—and initiating multi-level response protocols, including alerting mental health professionals or emergency services where necessary. By implementing such models within conversational platforms, we move toward a paradigm where AI not only supports mental wellness but also plays a life-saving role.

The proposed approach addresses a critical gap in current digital mental health interventions by embedding actionable intelligence at the user interface level, ensuring that flagging and escalation mechanisms are both immediate and contextually sensitive. Importantly, the system is designed to be ethically compliant and legally conscious, with safeguards to respect privacy and informed consent, while prioritizing user safety in crisis scenarios.

Moreover, the paper emphasizes the value of early intervention, advocating for the integration of such AI tools within school environments to normalize mental health conversations from a young age. By embedding support structures within educational systems, the long-term vision includes fostering resilience and emotional literacy, thereby reducing

the incidence of untreated psychological trauma.
While the model presents a robust framework, it remains subject to further validation, especially regarding false positives and the nuanced interpretation of linguistic cues. Future efforts must also consider socio-cultural variations in emotional expression and expand multilingual capabilities.

Ultimately, this work lays the foundation for an AI-driven, human-centered mental health safety net—one that blends technological precision with empathetic outreach to support individuals in their most vulnerable moments.

References

1. Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, 49(9), 1426-1448.
2. J. P. Pestian et al., "A machine learning approach to identifying suicidal adolescents in the emergency department," BMC Med. Inform. Decis. Mak., vol. 17, no. 1, p. 94, Aug. 2017.
3. Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical informatics insights, 10, 1178222618792860.
4. Inkster et al., "Machine learning and mental health: A systematic review of predictive models and an integrative framework," Psychol. Med., vol. 48, no. 9, pp. 1407–1422, Jun. 2018.
5. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. PLoS medicine, 15(11), e1002689.